# SocialNav-SUB: Benchmarking VLMs for Scene Understanding in Social Robot Navigation

**Michael J. Munje[1]*, Chen Tang[1], Shuijing Liu[1], Zichao Hu[1], Yifeng Zhu[1], Jiaxun Cui[1], Garrett Warnell[1,2], Joydeep Biswas[1], Peter Stone[1,3]**
[1]Department of Computer Science, The University of Texas at Austin
[2]Army Research Laboratory [3]Sony AI

**Abstract:** Robot navigation in dynamic, human-centered environments requires socially-compliant decisions grounded in robust scene understanding. Recent Vision-Language Models (VLMs) exhibit promising capabilities such as object recognition, common-sense reasoning, and contextual understanding—capabilities that align with the nuanced requirements of social robot navigation. However, it remains unclear whether VLMs can accurately understand complex social navigation scenes (e.g., inferring the spatial-temporal relations among agents and human intentions), which is essential for safe and socially compliant robot navigation. While some recent works have explored the use of VLMs in social robot navigation, no existing work systematically evaluates their ability to meet these necessary conditions. In this paper, we introduce the *Social Navigation Scene Understanding Benchmark (SocialNav-SUB)*, a Visual Question Answering (VQA) dataset and benchmark designed to evaluate VLMs for scene understanding in real-world social robot navigation scenarios. SocialNav-SUB provides a unified framework for evaluating VLMs against human and rule-based baselines across VQA tasks requiring spatial, spatiotemporal, and social reasoning in social robot navigation. Through experiments with state-of-the-art VLMs, we find that while the best-performing VLM achieves an encouraging probability of agreeing with human answers, it still underperforms simpler rule-based approach and human consensus baselines, indicating critical gaps in social scene understanding of current VLMs. Our benchmark sets the stage for further research on foundation models for social robot navigation, offering a framework to explore how VLMs can be tailored to meet real-world social robot navigation needs. An overview of this paper along with the code and data can be found at `https://larg.github.io/socialnav-sub`.

## 1 Introduction

Social robot navigation, defined as the ability for robots to move effectively and safely within human-populated environments while adhering to social norms, is a fundamental yet challenging task in robotics [1, 2]. As shown in Figure 1, navigating through social navigation scenarios requires robots to interpret human intentions, adhere to social norms, and reason about spatial and temporal interactions to respond to dynamic environments. While promising, learning-based methods that are trained on small datasets and conventional methods are often validated in controlled scenarios with a small number of people, thus falling short in handling the complexity and nuance in dynamic real-world social navigation scenarios [1, 3].

Recently, the research community has begun to explore whether advances in large Vision-Language Models (VLMs) can be leveraged as part of a solution to social robot navigation, as they have demonstrated strong capabilities in contextual understanding, commonsense reasoning, and chain-

*Correspondence to: `michaelmunje@utexas.edu`

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

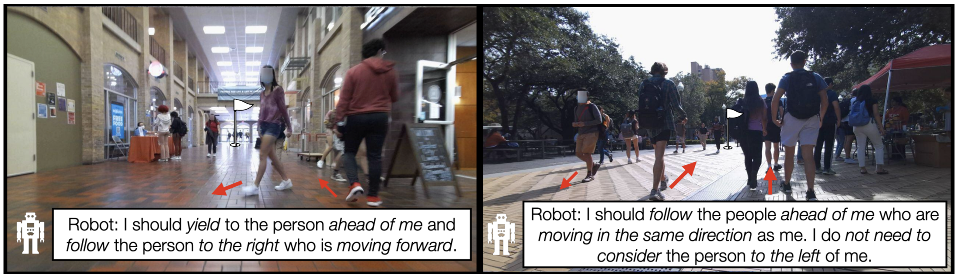

Robot: I should *yield* to the person *ahead of me* and *follow* the person *to the right* who is *moving forward*.
Robot: I should *follow* the people *ahead of me* who are *moving in the same direction* as me. I do *not need to consider* the person *to the left* of me.

Figure 1: **Examples of social robot navigation scenarios from SCAND [4].** The ability to determine socially compliant navigation actions requires understanding each dynamic scene by spatiotemporal reasoning (e.g. the movements of people in the scene) and social reasoning (inferring the navigation intentions of people in the scene).

of-thought reasoning [5, 6, 7]. Trained in diverse large-scale multimodal datasets that span various real-world scenarios, large VLMs often learn underlying patterns of human behavior that may implicitly encode an understanding of social norms [8]. However, in social navigation, the scene understanding capabilities of VLMs remains underexplored: Recent works like VLM-Social-Nav [9] have shown that using large VLMs for social robot navigation is promising, but their evaluations are limited to a small number of controlled scenarios and offer only preliminary insights. Moreover, studies such as SPACE [10] indicate that state-of-the-art large VLMs still lack robust spatial reasoning, raising questions about whether VLMs can understand scenes of complex, realistic social navigation scenarios at all or propose socially compliant actions for robots.

In light of these limitations, it remains essential to systematically evaluate whether VLMs can robustly handle what we consider as three critical dimensions of social robot navigation: **(1)** spatial reasoning [11], **(2)** spatiotemporal reasoning [12], and **(3)** the ability to interpret complex human intentions [13, 14]. Existing evaluations have offered only partial assessments [9, 10], often focusing on controlled settings or lacking temporal components, leading to an incomplete picture of how effectively large VLMs can infer human intentions and comply with social norms in realistic, dynamic scenarios. This gap underscores the need for a comprehensive benchmark that rigorously tests these capabilities and may guide the development of VLMs tailored to social robot navigation.

In this paper, we introduce the Social Navigation Scene Understanding Benchmark (SOCIALNAV-SUB), a novel Visual Question Answering (VQA) benchmark designed to evaluate VLMs on social robot navigation tasks. Shown in Figure 2, our benchmark utilizes data from a human-subject study conducted using social navigation scenarios from the SCAND dataset [4, 15], a robot social navigation dataset of socially compliant navigation demonstrations with dense crowds and diverse social settings. We use our comprehensive human-labeled VQA dataset to serve as ground-truth labels to systematically evaluate the performance of VLMs on scene understanding for social robot navigation for real-world scenarios. We run experiments on state-of-the-art large VLMs which reveal *notable performance gaps between state-of-the-art large VLMs and both human and rule-based baselines.*

SocialNav-SUB is a first-of-its-kind benchmark that enables roboticists to systematically evaluate and refine VLMs for real-world social robot navigation scenarios. By bridging the gap between VLM capabilities and the challenges of social robot navigation, our work provides a foundation for advancing the use of VLMs for social robot navigation. Our contributions are as follows:

1. **Social Navigation Scene Understanding Dataset:** We provide a human-labeled VQA dataset of 4968 unique questions and the accompanied human responses (serving as ground-truth labels) for social robot navigation tasks.
2. **Social Navigation VQA Benchmark for VLMs:** We introduce the first VQA benchmark for assessing VLMs' capabilities in social robot navigation scenarios using 60 unique scenarios from SCAND that evaluates agreement with human answers.
3. **Experiments using state-of-the-art large VLMs on our benchmark:** We evaluate several large VLMs (e.g., Gemini 2.0 and 2.5 [7], GPT-4o [6], OpenAI o4-mini [16], LLaVa-

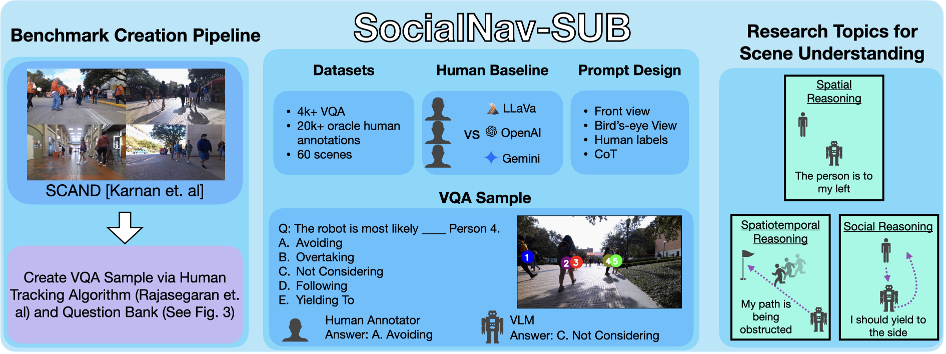


Figure 2: **An overview of SocialNav-SUB**, which facilitates the systematic evaluation of VLMs in social robot navigation scenarios. Using SCAND data, human-labeled VQA datasets, and various VLMs, this framework offers the evaluation of VLMs across multiple dimensions of scene understanding for social robot navigation that can enable advancements in prompt designs, social reasoning, and social robot navigation research in general.

Next-Video [17]) on our benchmark against human and rule-based baselines. All models perform worse than human oracle and rule-based performance.

## 2 Related Work

**VLMs in Robotics.** In robotics, VLMs have demonstrated considerable potential for various tasks such as robotic manipulation [18], task planning [19], and human-robot interaction [20, 21, 22]. The success of VLMs can be attributed to their ability to associate vision and language and generalize to unseen data in a zero-shot manner. For navigation, VLMs have been used for waypoint specification [18, 23], and instruction following [24, 25, 26]. However, these approaches often struggle in complex real-world environments, particularly in dynamic environments, due to limitations in VLMs' spatial reasoning capabilities [10, 27, 28]. This gap highlights the need for specialized evaluations and improvements of VLMs for tasks in dynamic environments, especially social navigation.

**Social Robot Navigation.** Early social robot navigation approaches relied on model-based techniques, such as the Social Force Model (SFM) [29] and proxemics-based methods [30], which used hand-engineered features to plan paths for robots. Learning-based techniques for social robot navigation, including Learning from Demonstration (LfD) [31, 4] and Reinforcement Learning (RL) [32, 33, 34, 35, 36], have shown promise in enabling robots to acquire and adapt socially compliant behaviors but are often trained on small and specialized data or simulations and struggle to generalize to complex dynamic scenarios. To address this, datasets for social robot navigation [4, 37] have been developed to provide more diverse and realistic social navigation scenarios, which can lead to improved generalization in social navigation models [38]. More recently, VQA datasets for social robot navigation have been explored [39], but are limited to qualitative evaluation and single images, when crucial information, such as a person's trajectory, may require a video representation. Fine-tuned VLMs have been explored for social robot navigation [9, 39], but are often evaluated in a limited number of simple, controlled scenarios. These scattered findings suggest that while VLMs *may* enhance social robot navigation, the specific capabilities that drive any observed improvements have yet to be clearly identified. Our work addresses this limitation by introducing a specialized benchmark to systematically evaluate whether VLMs can effectively perform spatial reasoning, spatiotemporal reasoning, and social reasoning for numerous social navigation scenarios.

**VQA Benchmarks for VLMs.** Recent years have seen the development of various VQA benchmarks to evaluate VLMs, assessing capabilities such as spatial reasoning [10], temporal reasoning for robot navigation [40, 41], scene understanding for autonomous driving [42, 43], and physical world comprehension [44]. While these benchmarks have advanced our understanding of VLMs' capabilities, they often lack specific focus on social robot navigation. Our work addresses this gap by introducing a specialized VQA benchmark for social robot navigation.

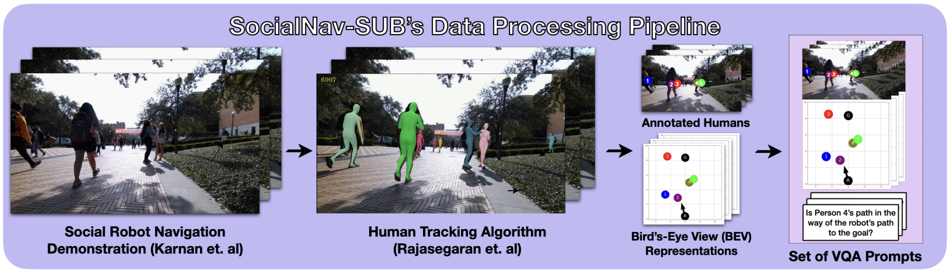

SocialNav-SUB's Data Processing Pipeline
6907
1
2
3
4
5
G
R
Annotated Humans
Social Robot Navigation Demonstration (Karnan et. al)
Human Tracking Algorithm (Rajasegaran et. al)
Bird's-Eye View (BEV) Representations
Is Person 4's path in the way of the robot's path to the goal?
Set of VQA Prompts

Figure 3: **The data processing pipeline for VQA prompts in SocialNav-SUB.** We first mine social robot navigation scenarios from SCAND [4], then use the PHALP algorithm [46] to provide human tracking and estimations of 3D locations, which are used to construct BEV representations of the scene and annotated images. Along with the annotated images and BEV representations, a set of carefully designed questions (more details in Appendix 7.7) that evaluate spatial reasoning, spatiotemporal reasoning, and social reasoning are used to provide VQA prompts.

## 3 SocialNav-SUB

To evaluate VLMs on scene understanding for social robot navigation, we present the **Social Navigation Scene Understanding Benchmark (SocialNav-SUB)**, a VQA benchmark for evaluating VLMs in social navigation scenarios. Following recent works that have demonstrated the effectiveness of visual grounding and object-centric representations [18, 45, 42], we provide numbered labels within visual markers for objects of relevance (in our case, pedestrians) for prompting and object-centric annotations; this provides the benchmarked VLMs clear visual references and contextually rich instructions. SocialNav-SUB is built on top of social navigation scenarios from SCAND that provide varying levels of crowd density and social navigation interactions and features the following: *Challenging social navigation scenarios* that capture the complexities of crowded and dynamic human environments; *Object-centric representations* combining both the robot's visual perspective and a bird's-eye view (BEV) containing pedestrian coordinate tracking for a richer object-centric representation; *A diverse question set* probing spatial reasoning, temporal understanding, and social reasoning; and *A robust human baseline*, where multiple annotators provide ground-truth responses for each scenario. All above features are expanded in the following subsections below.

### 3.1 Challenging Social Navigation Scenarios

To effectively evaluate VLMs' scene understanding capabilities in practical social robot navigation settings, we leverage the SCAND dataset [4] to construct SocialNav-SUB. SCAND features social robot navigation data collected by teleoperated mobile robots navigating in diverse and potentially crowded scenarios. In particular, we extract segments from SCAND that showcase moderate to high crowd density (average of 6.65 humans per scene, std. dev.: 2.80), close pedestrian proximity, and dynamically changing human motion. As illustrated in Figure 1, these densely occupied scenarios typically involve pedestrians that obstruct the robot's direct path to its goal. Hence, the teleoperated robots show complex, socially compliant interactions with the pedestrians, making these samples valuable for evaluating VLMs' scene understanding capabilities in real-world social navigation scenarios.

### 3.2 Rich and Object-Centric Visual Representations

The samples extracted from the SCAND dataset are in the form of RGB image sequences captured by the front-view camera mounted on the robot. While 2D image sequences may suffice for humans to infer the underlying spatial and social relations between the robots and pedestrians, state-of-the-art large VLMs are not necessarily good at extracting spatial or fine-grained object-level information from the same visual queries [10]. To mitigate this issue, some recent studies have shown that augmenting images with additional annotations (e.g., bounding boxes, color-coded labels) using off-the-shelf models can improve VLM performance in VQA tasks [18, 45].

Building on these visual prompting insights, we augment the original data samples with additional *object-centric* representations leveraging off-the-shelf vision models. As shown in Figure 3, we begin by employing the human tracking algorithm, PHALP [46], which tracks pedestrians and provides estimations of their 3D poses relative to the camera frame using monocular video input. Using the robot odometry data from SCAND, we transform the relative human poses at future timesteps into global poses relative to the robot pose in the initial frame, and apply Kalman smoothing to smooth the human poses. Afterwards, we use the camera intrinsics and extrinsics provided by SCAND to project the 3D coordinates of pedestrians into both front-view and BEV images. Finally, we annotate human positions in both views with numbered, color-coded circles. The resulting images with combined views preserve the original scene context while providing additional spatial and object-level information in a clear and structured format. In practical robotics stacks, such BEV representations can be constructed in real-time by either with learning-based methods [47] or by utilizing tracking cameras, depth sensors, and camera matrices to estimate global positions [48, 49]. Therefore, by querying VLMs with these enriched, object-centric visual inputs, SocialNav-SUB can provide practical insights into how to best leverage and complement state-of-the-art large VLMs for practical application in social robot navigation. To ensure fair comparisons between VLMs' outputs and human responses, the same set of visual inputs are provided to human annotators.

### 3.3 Diverse Scene Understanding Questions

Following the aforementioned data processing pipeline, we construct a set of samples consisting of multi-view image sequences with object-centric annotations, each representing a $2.5\,\mathrm{s}$ segment sampled at $4\,\mathrm{Hz}$. To comprehensively evaluate VLMs' scene understanding capabilities in social robot navigation, we design a set of multiple-choice questions (see Table 7.7 for more details and Appendix 7.5 for an example VQA prompt) that probe across three categories: 1) **Spatial reasoning:** Questions about describing the *spatial relations* in a *single frame*; 2) **Spatiotemporal reasoning:** Questions about describing the *motion* of the robot and pedestrians *over time*; and 3) **Social reasoning:** Questions that *infer whether* the robot and pedestrians are interacting and *how* they interact.

These categories of questions map onto what we see as key challenges of social robot navigation: perceiving spatial relations among participants (spatial reasoning), tracking their evolution as people move (spatiotemporal reasoning), and recognizing how humans and robots interact in social navigation (social reasoning). By evaluating VLMs across these dimensions, we gain a fine-grained understanding of where models excel or struggle in interpreting social navigation scenes.

### 3.4 Robust Human Baseline from Human-Subject Study

We conducted human-subject studies to collect human responses as ground-truth labels for these questions under an IRB-approved protocol. Given the subjective nature of many questions, particularly those related to social reasoning, we collected responses from at least five human participants for each scenario. Participants were recruited via Prolific [50] and were asked to complete a questionnaire containing questions for multiple randomly sampled scenarios.

By gathering this distribution of human responses, we can measure how closely each VLM output aligns with human judgments by computing the agreement between VLM answers and all human answers for a given question, which indicates the extent to which a model's performance approaches human-level responses. We define two metrics, **Probability of Agreement (PA)** and **Consensus-Weighted PA**, to measure how closely a set of answers (from a VLM, a particular human, or a rule-based baseline) aligns with human responses overall. Let $N_Q$ be the total number of questions; $N_H$ be the number of human respondents per question; $A_q$ be the evaluated answer (from a VLM or one human) to question $q$; and $A^h_{q,i}$ be the $i$-th human's answer for question $q$, where $i \in \{1, \dots, N_H\}$.

We define *Probability of Agreement (PA)* as the following:

$$\text{PA} \;=\; \frac{1}{N_Q} \sum_{q=1}^{N_Q} \Big( \frac{1}{N_H} \sum_{i=1}^{N_H} \mathbb{I}[\, A_q = A^h_{q,i} \,] \Big), \tag{1}$$

where $\mathbb{I}[\cdot]$ is an indicator function which outputs 1 if $A_q$ (the evaluated answer) exactly matches the $i$-th human's response $A^h_{q,i}$, and 0 otherwise for the corresponding multiple-choice question $q$. Summing over all human responses for each question yields the fraction of total (answer, human answer) pairs that agree. PA is essentially the expected cosine similarity between the model's predictions and the distribution of human responses. A higher PA indicates that the evaluated answers coincide more frequently with the collected human responses. We empirically found that it is common for humans to disagree on answers, indicating there is a degree of judgement involved for particular questions. This motivates a metric that can be more forgiving for subjective questions that humans disagree on and emphasize questions that have a strong consensus, to which we establish *Consensus-Weighted Probability of Agreement (CWPA)*. We start by defining

$$\mathrm{HA}_q \;=\; \max_{\alpha}\Big\{\frac{\text{\#(humans who answered } \alpha \text{ for question } q)}{N_H}\Big\},$$

i.e., $\mathrm{HA}_q$ is the fraction of human respondents that chose the most common answer $\alpha$ for question $q$. We then define:

$$\mathrm{CWPA} \;=\; \frac{1}{N_Q}\sum_{q=1}^{N_Q}\Big(\frac{1}{N_H\,\mathrm{HA}_q}\sum_{i=1}^{N_H}\mathbb{I}[\,A_q = A^h_{q,i}\,]\Big). \tag{2}$$

In this formulation, each agreement with a human response for question $q$ is scaled by $1/\mathrm{HA}_q$. Consequently, questions on which humans mostly concur (i.e., high $\mathrm{HA}_q$) impose a greater penalty for incorrect answers, while questions where humans are more divided have a lower penalty. This weighting ensures that VLM (or human) answers are held to a higher standard on "easier" questions where strong human agreement exists. Similar agreement-based metrics have been adopted to account for variability among human annotators when constructing VQA benchmarks [51]. Unlike their metrics, PA does not rely on a heuristically selected threshold to saturate the accuracy. CWPA further extends this by introducing a novel weighting scheme based on human consensus.

In addition to evaluation metrics, we utilize the human responses to construct two human baselines: An *Average Human Baseline*, which measures on average how often one human's response agrees with all other human responses and serves as an indicator of average human performance but may be susceptible to noise in responses from online human participants; and A *Human Oracle Baseline*, which selects the most common answer for each question from the human distribution and serves as a more robust estimate of expert-level human performance.

# 4 Empirical Results

Our central research question examines *how well state-of-the-art large VLMs that support image sequences capture spatial reasoning, scene understanding, and social reasoning in social robot navigation scenarios*. Focusing on this question, we aim to rigorously assess the capabilities and limitations of VLMs for understanding complex social robot navigation environments. We establish the benchmark with several representative models supporting video inputs across three categories: 1) **closed-source, general-purpose VLMs**, including GPT-4o [6] and Gemini 2.0 [7], which demonstrate strong overall performance in VQA tasks; 2) **reasoning VLMs**, including OpenAI o4-mini [16] and Gemini 2.5 [7], which are fine-tuned to enhance vision-language reasoning capabilities. While too computationally intensive for real-time deployment, they may be distilled into faster models [52] suitable for robotics applications; and 3) **open-source, deployable** VLMs, including LLaVa-Next-Video [17], which can run locally and are thus well-suited for robotics applications.

## 4.1 Experiment Process

Our experiment process begins by presenting survey prompts alongside their visual and BEV representations to the VLM, using the data processing pipeline previously shown in Figure 3. The format given to the VLMs closely resembles the same visual and text format that was received by human participants, ensuring fair comparison. Furthermore, we use chain-of-thought (CoT) reasoning as a prompting technique to carry out our experiments, since this is highly similar to the sequential manner in which humans provided answer labels, allowing for fair comparison. Specifically, our usage

Table 1: **Average Performance Across Question Categories.** The metrics used are PA and CWPA for all questions and for each question category, along with standard error across the questions. We highlight in bold the strongest VLM PA results, which may be statistically tied.

| Category | Model | All | | Spatial Reasoning | | Spatiotemporal Reasoning | | Social Reasoning | |
|---|---|---|---|---|---|---|---|---|---|
| | | PA | CWPA | PA | CWPA | PA | CWPA | PA | CWPA |
| **Baseline** | Human Oracle | 0.74 ± 0.00 | 1.0 ± 0.00 | 0.71 ± 0.01 | 1.0 ± 0.00 | 0.73 ± 0.01 | 1.0 ± 0.00 | 0.76 ± 0.01 | 1.0 ± 0.00 |
| | Average Human | 0.60 ± 0.00 | 0.80 ± 0.00 | 0.56 ± 0.01 | 0.79 ± 0.00 | 0.59 ± 0.01 | 0.80 ± 0.00 | 0.62 ± 0.00 | 0.81 ± 0.00 |
| | Rule-Based | 0.64 ± 0.00 | 0.84 ± 0.00 | 0.57 ± 0.01 | 0.79 ± 0.01 | 0.62 ± 0.01 | 0.84 ± 0.01 | 0.71 ± 0.00 | 0.92 ± 0.00 |
| **VLM** | Gemini 2.0 | 0.58 ± 0.00 | 0.79 ± 0.00 | **0.55 ± 0.01** | 0.77 ± 0.01 | 0.46 ± 0.01 | 0.64 ± 0.01 | 0.63 ± 0.01 | 0.84 ± 0.01 |
| | Gemini 2.5 | 0.54 ± 0.00 | 0.73 ± 0.01 | 0.51 ± 0.01 | 0.72 ± 0.01 | 0.52 ± 0.01 | 0.73 ± 0.01 | 0.55 ± 0.01 | 0.73 ± 0.01 |
| | GPT-4o | 0.50 ± 0.00 | 0.69 ± 0.01 | **0.56 ± 0.01** | 0.79 ± 0.01 | 0.51 ± 0.01 | 0.71 ± 0.01 | 0.47 ± 0.01 | 0.63 ± 0.01 |
| | o4-mini | **0.62 ± 0.01** | 0.82 ± 0.01 | 0.54 ± 0.01 | 0.74 ± 0.01 | **0.59 ± 0.01** | 0.79 ± 0.01 | **0.66 ± 0.01** | 0.87 ± 0.01 |
| | LLaVa-Next-Video | 0.46 ± 0.01 | 0.61 ± 0.01 | 0.35 ± 0.01 | 0.46 ± 0.01 | **0.58 ± 0.01** | 0.79 ± 0.01 | 0.48 ± 0.01 | 0.62 ± 0.01 |

of CoT provides the previous answers of the VLM for future questions which may help it deduce the answer to question; for example, the pedestrian is at the left in the beginning and the end and the goal is on the right, so the pedestrian is likely not obstructing the path to the goal. The responses generated by the VLM are then compared against human responses from the human dataset using the PA and CWPA metrics, previously defined in Equations 1 and 2

Humans can naturally infer the underlying spatial and social relations between the robots and pedestrians, making them excellent reference points of performance. On the other hand, are large VLMs truly necessary for analyzing these social robot navigation scenarios, or can a simpler, rule-based system suffice? To address both of these, we utilize the two human baselines previously defined in Section 3.4, the *Human Oracle Baseline* and the *Average Human Baseline*, as well as a *Rule-Based Baseline*, which uses the position data of pedestrians in the scene and uses a set of hand-crafted rules to generate answers to VQA prompts (for more details, see Appendix 7.10).

### 4.2 Benchmarking Results

We run our experiments by querying each VLM model once per unique question using default hyperparameters for each VLM. The average results over all questions and question categories is shown in Table 1. Among the models evaluated, OpenAI o4-mini achieves the highest overall performance, but still has a considerable gap compared to the human oracle and rule-based baselines. This performance gap suggests that state-of-the-art large VLMs are not yet fully ready for the challenges of scene understanding for social robot navigation.

When examining performance across the three question categories, models consistently lag behind the human oracle and the rule-based baseline, though the extent of the gap varies by category and perform up to par with the average human baseline. In spatial reasoning, the consensus among humans (human oracle) far exceeds that of the best models, indicating that current large VLMs struggle to accurately interpret spatial relationships compared to human observers. A similar finding is observed in spatiotemporal reasoning, where models show greater difficulty at capturing dynamic changes over time. In contrast, in social reasoning tasks, models perform relatively closer to human oracle levels and can even slightly outperform the average human baseline, suggesting that large VLMs are somewhat more adept at interpreting social cues and interactions than they are at understanding spatial relationships, although there remains a noticeable gap. Empirically, we found many cases of VLMs failing on questions with high human consensus in all three reasoning categories, especially in cases of high crowd densities, we provide qualitative examples within Appendix 7.6.

### 4.3 Discussion

Overall, our evaluation reveals that while state-of-the-art large VLMs like OpenAI o4-mini and Gemini 2.0 show promising advances, they still fall short of human oracle and rule-based performance across key reasoning tasks. Although models come closer to human oracle performance in social reasoning tasks, the results suggest that significant improvements are needed before these large VLMs can reliably support complex, real-world social robot navigation.

Table 2: **Ablation experiment of querying strategies**. The metric used is Probability of Agreement (PA). The baseline row BEV+CoT represents the performance with both CoT and BEV prompts enabled. The subsequent rows show the effects of removing either CoT or BEV components.

| Model | Ablation | Spatial Reasoning | Spatiotemporal Reasoning | Social Reasoning |
|---|---|---|---|---|
| GPT-4o | CoT+BEV | 0.56 ± 0.01 | 0.51 ± 0.01 | 0.47 ± 0.01 |
| | No CoT | 0.58 ± 0.01 | 0.53 ± 0.01 | 0.35 ± 0.01 |
| | No BEV | 0.51 ± 0.01 | 0.44 ± 0.01 | 0.42 ± 0.01 |
| LLaVa-Next-Video | CoT+BEV | 0.35 ± 0.01 | 0.58 ± 0.01 | 0.48 ± 0.01 |
| | No CoT | 0.35 ± 0.01 | 0.58 ± 0.01 | 0.38 ± 0.01 |
| | No BEV | 0.35 ± 0.01 | 0.61 ± 0.01 | 0.46 ± 0.01 |
| Gemini 2.0 | CoT+BEV | 0.55 ± 0.01 | 0.46 ± 0.01 | 0.63 ± 0.01 |
| | No CoT | 0.56 ± 0.01 | 0.48 ± 0.01 | 0.58 ± 0.01 |
| | No BEV | 0.56 ± 0.01 | 0.46 ± 0.01 | 0.64 ± 0.01 |

We also performed a series of ablation experiments to study the impact of querying strategies to the model performance. The results are summarized in Table 13 (more details in Appendix 7.9). Our first ablation experiment analyzed the impact of CoT reasoning and found that it significantly enhances social reasoning performance for all models, likely due to the structured inference it provides for complex tasks. We also performed another ablation experiment investigating the impact of BEV scene representations and found that some models may benefit significantly, while other models show minimal changes. This suggests that BEV effectiveness depends on the VLM, but can be validated through SOCIALNAV-SUB. A further ablation experiment looked at the effectiveness of better spatial and spatiotemporal reasoning capabilities and found stronger performance on social reasoning questions, suggesting that current VLMs are limited by spatial reasoning capabilities but may be improved with fine-tuning on spatial reasoning data [53, 27] while maintaining performance on higher-level scene understanding. These experiments highlight the usefulness of SOCIALNAV-SUB in informing how VLMs can be best utilized and further improved for social robot navigation.

Finally, we revisited our original assumption described in Section 1 that accurate scene understanding is a prerequisite for the practical usage of VLMs in real-world navigation tasks. To validate this claim, we carried out an experiment to examine the impact of scene understanding on the task of waypoint selection (see Appendix 7.1). Results indicated that providing additional scene context improved the alignment of answers with those chosen by human operators across all models, especially for reasoning models. These findings reinforce the value of our SOCIALNAV-SUB benchmark in advancing VLMs for real-world social robot navigation tasks.

## 5 Conclusion

This paper introduced the Social Navigation Scene Understanding Benchmark (SOCIALNAV-SUB), a novel VQA benchmark designed to evaluate VLMs within complex social robot navigation scenarios. Drawing on crowded and dynamic environments from the SCAND dataset, SOCIALNAV-SUB provides object-centric visual representations, including augmented front-view images and BEV prompts, paired with a diverse set of questions targeting spatial, spatiotemporal, and social reasoning. By grounding these evaluations with a human-subject study, the benchmark offers clear, quantifiable metrics that reflect human-like understanding and decision-making in social navigation. SOCIALNAV-SUB advances the state of the art by highlighting specific strengths and weaknesses of current VLMs in handling intricate social scenes, thereby setting a clear agenda for future research. It enables researchers to systematically compare models, refine prompting strategies, and develop new methods to bridge the gap between machine and human understanding of social navigation scenes and allows for the iterative improvement of VLMs in real-world applications, ultimately guiding the development of more socially aware and reliable robotic systems.

## 6 Limitations and Future Work

While SocialNav-SUB advances the evaluation of VLMs for social robot navigation, it has two limitations. First, the benchmark currently relies on scenarios from the SCAND dataset, which, despite the diverse scenarios and dense crowds (examples can be seen in the Appendix), is limited to social navigation in a university campus setting. Second, while initial experiments provide valuable insights, they are based on a limited set of models and scenarios; further exploration with a broader range of large VLMs, datasets, and refined methodologies is necessary to overcome these challenges and enhance the benchmark's applicability.

Looking ahead, several promising avenues can further enhance and leverage the capabilities of SocialNav-SUB. First, expanding the dataset to include additional social robot navigation datasets could expand its diversity and robustness, offering a more comprehensive evaluation of model capabilities. Additionally, fine-tuning VLMs on the human dataset provided in SocialNav-SUB may lead to VLMs that are more capable of social robot navigation. Another promising avenue is expanding upon the VLM models evaluated; some VLMs of interest include VLMs fine-tuned for spatial reasoning and VLMs fine-tuned for social robot navigation. Lastly, an interesting future direction is evaluating hybrid approaches that utilize VLMs in specific ways (such as social reasoning) while having dedicated modules to cover their weaknesses. By offering a targeted evaluation framework across multiple reasoning categories, SocialNav-SUB can not only systematically evaluate VLM performance and highlight weaknesses but also guide future improvements in VLMs for both scene understanding and socially compliant navigation, enabling the development of more reliable real-world robotics systems. Since we will open-source SocialNav-SUB and plan to reliably maintain it, much of the infrastructure and support to pursue these future endeavors will be readily available.

### Acknowledgments

This work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-24-1-2550), ARO (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork) Lockheed Martin, and UT Austin's Good Systems grand challenge. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research. We thank Georgios Pavlakos for helpful discussions on PHALP-based 3D position estimation and Haresh Karnan for clarifications on the SCAND dataset.

## References

[1] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld, and J. Oh. Core challenges of social robot navigation: A survey. *J. Hum.-Robot Interact.*, 12(3), Apr. 2023. doi:10.1145/3583741. URL https://doi.org/10.1145/3583741.

[2] A. Francis, C. Pérez-D'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, H.-T. L. Chiang, M. Everett, S. Ha, J. Hart, J. P. How, H. Karnan, T.-W. E. Lee, L. J. Manso, R. Mirksy, S. Pirk, P. T. Singamaneni, P. Stone, A. V. Taylor, P. Trautman, N. Tsoi, M. Vázquez, X. Xiao, P. Xu, N. Yokoyama, A. Toshev, and R. Martín-Martín. Principles and guidelines for evaluating social robot navigation algorithms, 2023. URL https://arxiv.org/abs/2306.16740.

[3] A. H. Raj, Z. Hu, H. Karnan, R. Chandra, A. Payandeh, L. Mao, P. Stone, J. Biswas, and X. Xiao. Rethinking social robot navigation: Leveraging the best of two worlds. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16330–16337, 2024. doi:10.1109/ICRA57147.2024.10611710.

[4] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation, 2022. URL https://arxiv.org/abs/2203.15041.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

[6] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, and A. C. et. al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[7] G. Team and P. G. et. al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[8] Z. Hu, Y. Ren, J. Li, and Y. Yin. Viva: A benchmark for vision-grounded decision-making with human values, 2024. URL https://arxiv.org/abs/2407.03000.

[9] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha. Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models, 2024. URL https://arxiv.org/abs/2404.00210.

[10] S. K. Ramakrishnan, E. Wijmans, P. Kraehenbuehl, and V. Koltun. Does spatial cognition emerge in frontier models?, 2024. URL https://arxiv.org/abs/2410.06468.

[11] F. Kessler, J. Frankenstein, and C. A. Rothkopf. Human navigation strategies and their errors result from dynamic interactions of spatial uncertainties. *Nature Communications*, 15(1):5677, 2024.

[12] A. D. Ekstrom and P. F. Hill. Spatial navigation and memory: A review of the similarities and differences relevant to brain models and age. *Neuron*, 111(7):1037–1049, 2023. ISSN 0896-6273. doi:https://doi.org/10.1016/j.neuron.2023.03.001. URL https://www.sciencedirect.com/science/article/pii/S0896627323001691.

[13] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz. The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS one*, 5(4):e10047, 2010.

[14] M. Moussaïd, D. Helbing, S. Garnier, A. Johansson, M. Combe, and G. Theraulaz. Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proceedings of the Royal Society B: Biological Sciences*, 276(1668):2755–2762, 2009.

[15] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone. Socially Compliant Navigation Dataset (SCAND), 2022. URL https://doi.org/10.18738/T8/0PRYRH.

[16] OpenAI. Openai o3 and o4-mini system card, 2025. URL https://openai.com/index/o3-o4-mini-system-card/.

[17] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[18] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, Q. Vuong, T. Zhang, T.-W. E. Lee, K.-H. Lee, P. Xu, S. Kirmani, Y. Zhu, A. Zeng, K. Hausman, N. Heess, C. Finn, S. Levine, and B. Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL https://arxiv.org/abs/2402.07872.

[19] Z. Yang, C. Garrett, D. Fox, T. Lozano-Pérez, and L. P. Kaelbling. Guiding long-horizon task and motion planning with vision language models. *arXiv preprint arXiv:2410.02193*, 2024.

[20] P. Chang, S. Liu, and K. Driggs-Campbell. Learning visual-audio representations for voice-controlled robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[21] P. Chang, S. Liu, T. Ji, N. Chakraborty, K. Hong, and K. R. Driggs-Campbell. A data-efficient visual-audio representation with intuitive fine-tuning for voice-controlled robots. In *Conference on Robot Learning (CoRL)*, 2023.

[22] Z. Dong, W. Zhang, X. Huang, H. Ji, X. Zhan, and J. Chen. Hubo-vlm: Unified vision-language model designed for human robot interaction tasks. *arXiv preprint arXiv:2308.12537*, 2023.

[23] A. J. Sathyamoorthy, K. Weerakoon, M. Elnoor, A. Zore, B. Ichter, F. Xia, J. Tan, W. Yu, and D. Manocha. Convoi: Context-aware navigation using vision language models in outdoor and indoor environments, 2024. URL https://arxiv.org/abs/2403.15637.

[24] H.-T. L. Chiang, Z. Xu, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, F. Xia, J. Hsu, J. Hoech, P. Florence, S. Kirmani, S. Singh, V. Sindhwani, C. Parada, C. Finn, P. Xu, S. Levine, and J. Tan. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs, 2024. URL https://arxiv.org/abs/2407.07775.

[25] K. Weerakoon, M. Elnoor, G. Seneviratne, V. Rajagopal, S. H. Arul, J. Liang, M. K. M. Jaffar, and D. Manocha. Behav: Behavioral rule guided autonomy using vlms for robot navigation in outdoor scenes, 2024. URL https://arxiv.org/abs/2409.16484.

[26] N. Hirose, C. Glossop, A. Sridhar, D. Shah, O. Mees, and S. Levine. Lelan: Learning a language-conditioned navigation policy from in-the-wild videos, 2024. URL https://arxiv.org/abs/2410.03603.

[27] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. Guibas, and F. Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

[28] Y. Tang, A. Qu, Z. Wang, D. Zhuang, Z. Wu, W. Ma, S. Wang, Y. Zheng, Z. Zhao, and J. Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning, 2024. URL https://arxiv.org/abs/2410.16162.

[29] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51 (5):4282, 1995.

[30] J. Mumm and B. Mutlu. Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 331–338, 2011. doi:10.1145/1957656.1957786.

[31] N. Hirose, D. Shah, A. Sridhar, and S. Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 2023.

[32] K. Zhu and T. Zhang. Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, 26(5):674–691, 2021.

[33] Y. F. Chen, M. Everett, M. Liu, and J. P. How. Socially aware motion planning with deep reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1343–1350, 2017.

[34] C. Chen, Y. Liu, S. Kreiss, and A. Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6015–6022, 2019.

[35] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. L. McPherson, J. Geng, and K. Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 12015–12021, 2023.

[36] S. Liu, H. Xia, F. C. Pouria, K. Hong, N. Chakraborty, Z. Hu, J. Biswas, and K. Driggs-Campbell. Height: Heterogeneous interaction graph transformer for robot navigation in crowded and constrained environments. *arXiv preprint arXiv:2411.12150*, 2025. URL https://arxiv.org/abs/2411.12150.

[37] D. M. Nguyen, M. Nazeri, A. Payandeh, A. Datar, and X. Xiao. Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7442–7447. IEEE, 2023.

[38] N. Hirose, D. Shah, A. Sridhar, and S. Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2024. doi:10.1109/LRA.2023.3329626.

[39] A. Payandeh, D. Song, M. Nazeri, J. Liang, P. Mukherjee, A. H. Raj, Y. Kong, D. Manocha, and X. Xiao. Social-llava: Enhancing robot navigation through human-language reasoning in social spaces. *arXiv preprint arXiv:2501.09024*, 2024.

[40] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang. Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation. *arXiv preprint arXiv:2409.13682*, 2024.

[41] A. Zhang, C. Eranki, C. Zhang, J.-H. Park, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, A. Bagad, M. Esteva, and J. Biswas. Towards robust robot 3d perception in urban environments: The ut campus object dataset, 2023. URL https://arxiv.org/abs/2309.13549.

[42] W. Wang, C. Duan, Z. Peng, Y. Liu, and B. Zhou. Embodied scene understanding for vision language models via metavqa. *arXiv preprint arXiv:2501.09167*, 2025.

[43] S. Sreeram, T.-H. Wang, A. Maalouf, G. Rosman, S. Karaman, and D. Rus. Probing multimodal llms as world models for driving. *arXiv preprint arXiv:2405.05956*, 2024.

[44] W. Chow*, J. Mao*, B. Li, D. Seita, V. C. Guizilini, and Y. Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025.

[45] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[46] J. Rajasegaran, G. Pavlakos, A. Kanazawa, and J. Malik. Tracking people by predicting 3d appearance, location and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2740–2749, 2022.

[47] A. Zhang, H. Sikchi, A. Zhang, and J. Biswas. Creste: Scalable mapless navigation with internet scale priors and counterfactual guidance. *arXiv preprint arXiv:2503.03921*, 2025.

[48] V. Tadic, A. Toth, Z. Vizvari, M. Klincsik, Z. Sari, P. Sarcevic, J. Sarosi, and I. Biro. Perspectives of realsense and zed depth sensors for robotic vision applications. *Machines*, 10(3):183, 2022.

[49] N. Aharony, A. Meshurer, M. Krakovski, Y. Parmet, I. Melzer, and Y. Edan. Comparative analysis of cameras and software tools for skeleton tracking. *IEEE Sensors Journal*, 2024.

[50] Prolific. Prolific. https://www.prolific.com, 2014. *Accessed on [date accessed]*.

[51] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[52] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

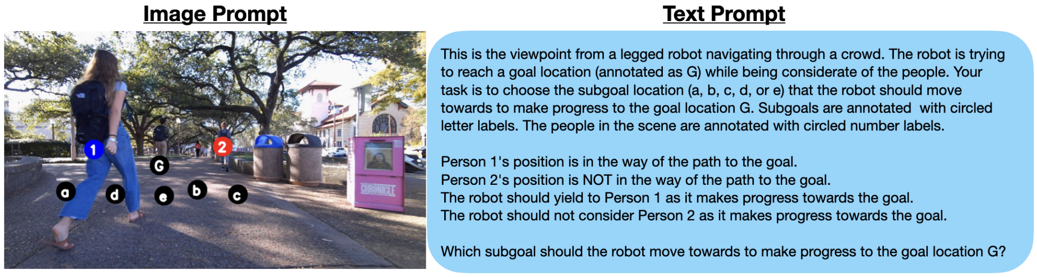
[53] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu. Spatialrgpt: Grounded spatial reasoning in vision language models, 2024. URL https://arxiv.org/abs/2406.01584.

# 7 Appendix

## 7.1 Waypoint Selection Experiments

To further demonstrate the practical value of SocialNav-SUB in real-world social robot navigation, we conduct preliminary experiments examining how scene understanding influences VLMs' performance in waypoint selection [18, 45]. Specifically, given visual observations and a set of candidate future waypoints annotated on the images, we prompt VLMs to select the waypoint that makes progress towards the goal while being considerate of the humans in the scene (see Figure 4). One of the candidate waypoints corresponds to the ground-truth future position of the robot as determined by the human operator. We evaluate the VLMs by comparing their selections to those made by the human operator. In addition to visual input, we incorporate scene context from various sources into the text prompts to assess their impact on waypoint selection. In this preliminary study, we condition the prompts on spatial reasoning and social reasoning context derived from predicted interactions among agents in the scene. These are provided in the form of answers to the **Person End Goal Obstruction** and **Robot Action to Person at End**.

The experimental results are presented in Table 3. Overall, when scene context is extracted from the human oracle's responses, VLM performance significantly improves compared to using no context or randomly generated context, and also shows slight improvement over using scene context predicted by the model itself. While preliminary, these findings suggest that accurate social scene context helps VLMs infer the ground-truth waypoints more effectively. This implies that enhancing a VLM's scene understanding capabilities can enable it to more accurately interpret social context and subsequently select appropriate navigation actions, thereby improving overall navigation performance. Our SocialNav-SUB benchmark provides the community with a valuable dataset and evaluation toolkit to support exploration along this direction.



This is the viewpoint from a legged robot navigating through a crowd. The robot is trying to reach a goal location (annotated as G) while being considerate of the people. Your task is to choose the subgoal location (a, b, c, d, or e) that the robot should move towards to make progress to the goal location G. Subgoals are annotated with circled letter labels. The people in the scene are annotated with circled number labels.

Person 1's position is in the way of the path to the goal.
Person 2's position is NOT in the way of the path to the goal.
The robot should yield to Person 1 as it makes progress towards the goal.
The robot should not consider Person 2 as it makes progress towards the goal.

Which subgoal should the robot move towards to make progress to the goal location G?

Figure 4: **An example of the waypoint selection VQA task.** This particular example highlights using scene context from the human oracle. Having no context removes the middle portion of the text prompt that includes the context, and having random context randomizes each relational action for the context (such as "avoiding").

Table 3: Accuracy of various VLMs in selecting the same waypoint as the human operator under social scene contexts from different sources: a random generator, the model itself, or the consensus from human participants (i.e., human oracle). The evaluation results are averaged over 5 runs, and we report mean accuracy ± standard error.

| Model | No Context | Random | Same-Model | Human Oracle |
|---|---|---|---|---|
| o4-mini | 36.14% ± 1.31% | 30.88% ± 2.12% | 38.95% ± 1.51% | 46.32% ± 1.19% |
| Gemini 2.0 | 37.19% ± 4.75% | 34.74% ± 4.09% | 41.05% ± 2.58% | 46.67% ± 3.62% |
| Gemini 2.5 | 34.39% ± 1.97% | 32.28% ± 1.72% | 37.19% ± 1.70% | 42.11% ± 2.88% |

## 7.2 Selecting Challenging Scenes for SocialNav-SUB

We curated 60 challenging scenes from SCAND to construct SocialNav-SUB. Candidate scenarios were ranked using a weighted linear score over features we hypothesized to correlate with difficulty for social robot navigation: (i) crowd size, (ii) the number of people within close proximity to the robot, and (iii) the robot's lateral (left–right) movement. We computed a weighted sum of these features and selected top-scoring scenes. The resulting set of scenarios spans across various environment types (e.g., outdoor walkways, narrow doorways/corridors, sidewalks, and street crossings) and a wide range of crowd densities (i.e., 1–13 pedestrians with mean = 6.65, SD = 2.80). Figure 5 shows four representative examples.

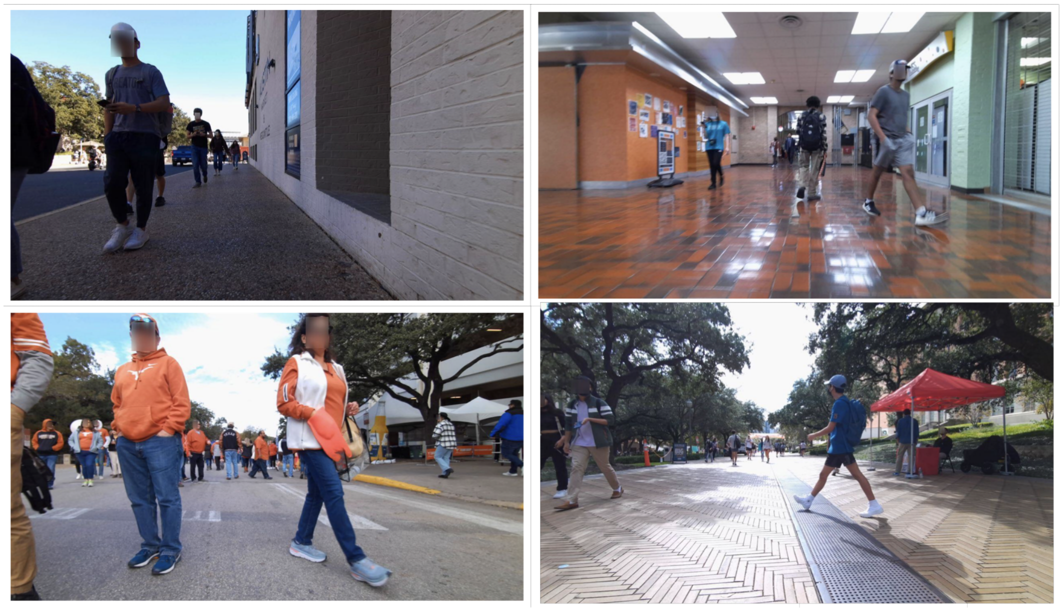


Figure 5: **Examples of scenes from SOCIALNAV-SUB.** These illustrate variation in environment type, crowd density, and human–robot proximity. SOCIALNAV-SUB comprises 60 social robot navigation scenarios in total.

### 7.3 Validation of 3D Pose Estimation Pipeline

As described in Section 3.2, we estimate 3D human pose trajectories from videos using PHALP and apply Kalman smoothing to filter the estimated trajectories. Since SCAND does not provide 3D human pose labels, we validated this pipeline and tuned the hyperparameters on the CODa dataset [41], which provides high-quality labels 3D human pose annotations derived from human-annotated 3D bounding boxes and human-in-the-loop SLAM-based localization. We tuned the Kalman smoothing hyperparameters on CODa by minimizing a weighted sum of average displacement error and angular displacement error over trajectories across multiple scenarios. The resulting hyperparameters are then used in the SOCIALNAV-SUB pipeline. Figure 6 shows a CODa scenario with our estimates and the labels provided by CODa. Estimates achieve an average displacement error of $0.67 \pm 0.14\,\mathrm{m}$ across all samples. Empirically we have observed errors are lower for well-observed pedestrians and larger under heavy occlusion.

### 7.4 Human-Subject Study Details

As mentioned in Section 3.4, we conducted a human-subject study under an IRB-approved protocol to collect human data to establish an evaluation method for SOCIALNAV-SUB. We conducted our human-subject study using Prolific [50] with 153 participants that were randomly selected across the U.S whose age's range from 18 to 80 (avg. 37.70, std. dev 13.40) with gender ratios of 44% male, 54% female, and 2% other. Figure 7 shows an example of the interface the humans were provided for the human-subject study. Humans sequentially answered questions for each scenario

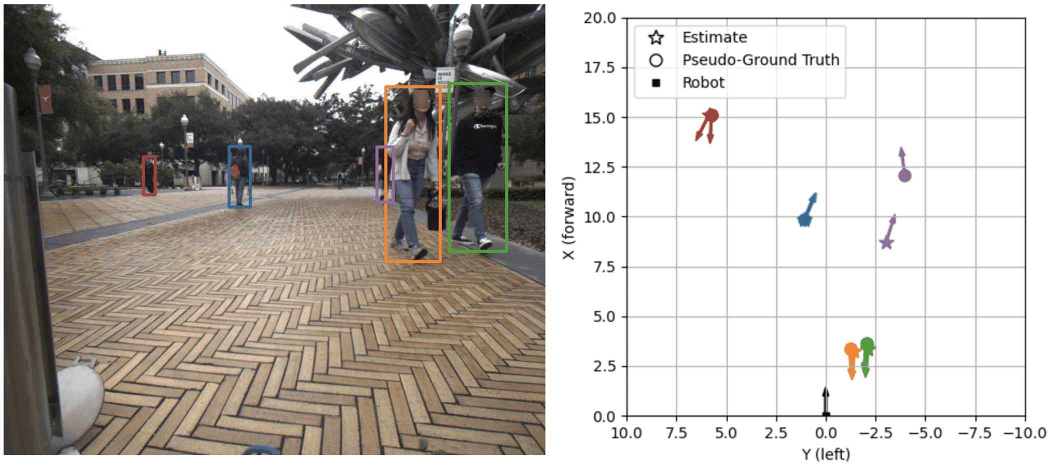
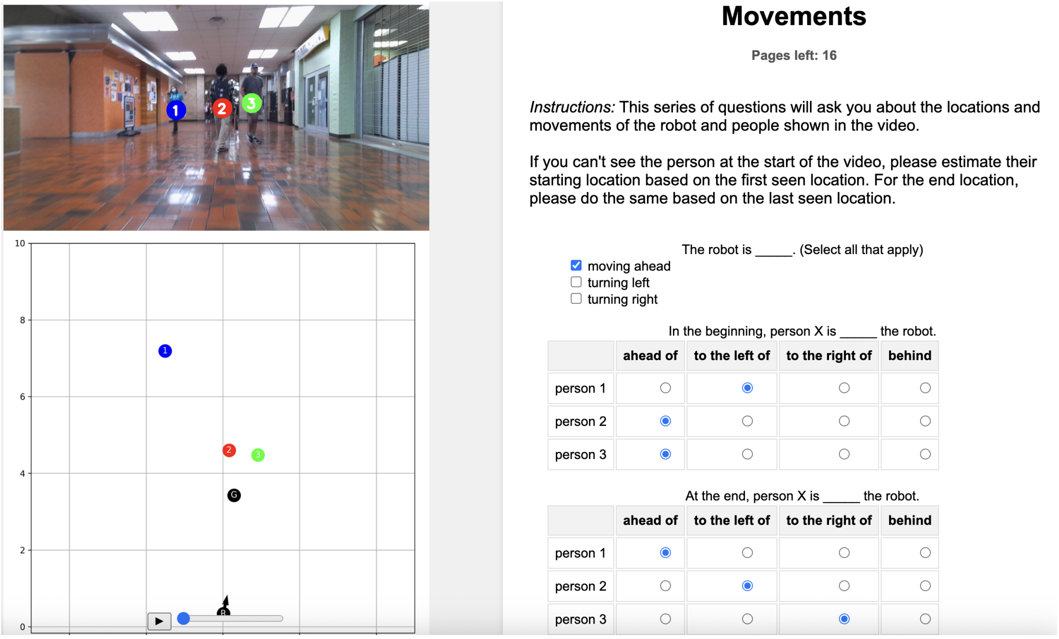


Figure 6: **CODa example for 3D pose pipeline validation. Left:** input image with PHALP bounding box detections. **Right:** BEV positions and headings after Kalman smoothing. Estimates are generally close ($< 1$ m displacement error) to pseudo–ground truth for well-observed pedestrians; errors increase for heavily occluded subjects.

in the following order: spatial reasoning questions, spatiotemporal reasoning questions, and social reasoning questions.



**Movements**

Pages left: 16

*Instructions:* This series of questions will ask you about the locations and movements of the robot and people shown in the video.

If you can't see the person at the start of the video, please estimate their starting location based on the first seen location. For the end location, please do the same based on the last seen location.

The robot is _____. (Select all that apply)

☑ moving ahead
☐ turning left
☐ turning right

In the beginning, person X is _____ the robot.

| | ahead of | to the left of | to the right of | behind |
|---|---|---|---|---|
| person 1 | ○ | ● | ○ | ○ |
| person 2 | ● | ○ | ○ | ○ |
| person 3 | ● | ○ | ○ | ○ |

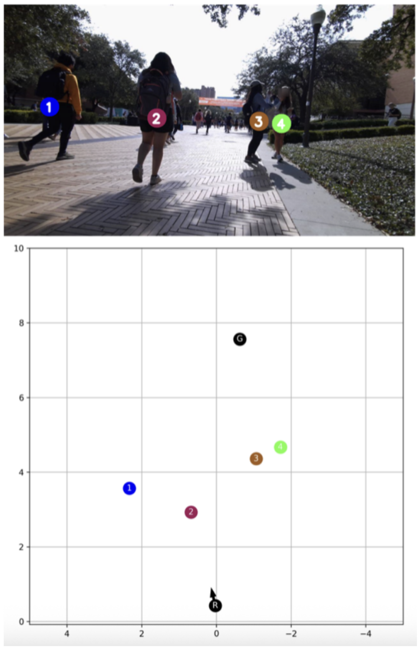
At the end, person X is _____ the robot.

| | ahead of | to the left of | to the right of | behind |
|---|---|---|---|---|
| person 1 | ● | ○ | ○ | ○ |
| person 2 | ○ | ● | ○ | ○ |
| person 3 | ○ | ○ | ● | ○ |

Figure 7: **An example of a survey page shown to human participants.** Prior to answering the survey questions, human subjects were given human-subject study participation instructions, requirements, and instructions about the survey content.

### 7.5 VQA Prompt Details

To provide fair comparison between humans and VLMs, we provided VLMs with highly similar input. In Figure 8, we provide a full VQA example of what the VLM receives as input. Chain-of-thought reasoning was used in the main experiments outlined in Section 4.3 and this particular usage consisted of sequentially asking the VLM questions, where later questions require higher-level reasoning, and providing the VLM its answers to the relevant questions within the prompt.



Title: Navigating Through Crowds Survey
This is a research survey about walking through crowds. You will be shown sequences of images of a robot navigating around people. The survey consists of multiple choice questions about the behavior of people and robots in a shared space.

Survey Description & Instructions:
The sequences of images are from the robot's perspective. Some visible people in the sequences of images have a unique circled number. The robot is generally moving forward but may turn or change its speed. We will ask you questions about these different sequences of images representing different scenarios.

Instructions: This series of questions will ask you about the locations and movements of the robot and people shown in the sequence of images.

If you can't see the person at the start of the video, please estimate their starting location based on the first seen location. For the end location, please do the same based on the last seen location.

In the beginning, Person 2 is _____ the robot.
Possible answers: “ahead of”, “to the left of”, “to the right of”, “behind”

Please provide the answer to the single question in JSON format (NOT a list) as follows:
{"answer": "<one of the possible answers>"}

Ensure the response is in JSON format and includes only one key, “answer"

Figure 8: **An example of a full VQA prompt shown to VLMs.** This context closely resembles the instructions that were provided to human participants for the human-subject study. In addition to the image shown on the left, the VLM also receives the next 9 images in the sequence.

### 7.6 Failure Case Analysis

As mentioned in Section 4.2, we found cases of VLMs in the experiment failing on questions with high human consensus in all reasoning categories, especially in cases of high crowd densities; we show these failure cases in Figure 9. We also highlight cases where VLMs can provide success, shown in Figure 10. These cases were automatically selected based on the entropy of the VLM answers and human answers.

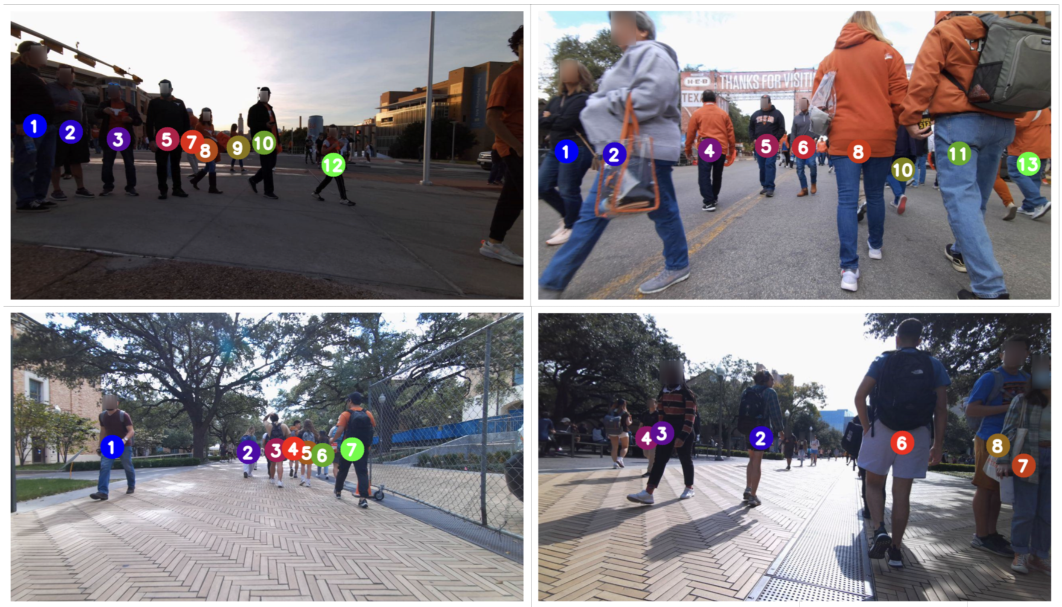


Figure 9: **Examples of failure cases for VLMs.** *Top-left:* Failing to recognize that person 5 is on the left. *Top-right:* Failing to recognize that person 4 ends up further away. *Bottom-left:* Answering that the distant person 3 should be avoided. *Bottom-right:* Incorrectly answering that an action should be taken with respect to person 7, although all humans did not think they were relevant. These examples were selected automatically based upon the entropy of the VLM answers and human answers.

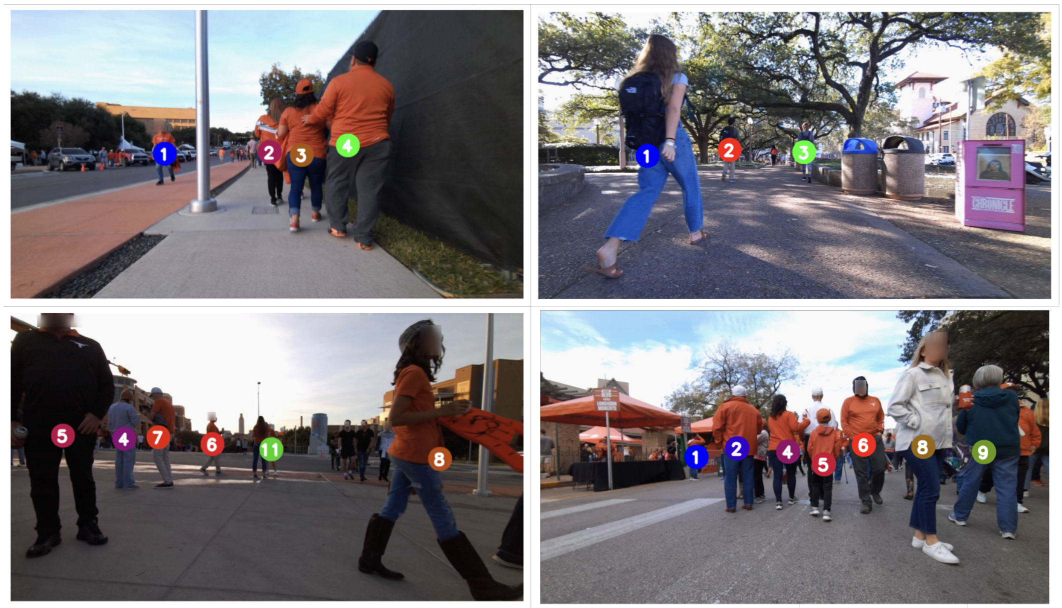


Figure 10: **Examples of success cases for VLMs.** *Top-left:* All VLMs correctly infer that person 1 is not obstructing the path to the goal. *Top-right:* Gemini correctly predicts that person 1 should be avoided (the other VLMs incorrectly predict this). *Bottom-left:* GPT-4o correctly answers that person 5 is on the left, whereas both Gemini and LLaVa-Next-Video answer that person 5 is behind the robot. *Bottom-right:* Most VLMs (but not all) predict that person 6 is being considered as the robot is moving towards the goal, similar to the distribution among human answers. These examples were automatically selected based upon the entropy of the VLM answers and the human answers.

**Quantitative summary.** We summarize quantitative findings shown in Tables 4 and 5 for VLM failure cases (defined as the model's chosen answer received zero human probability); we report failure rates in percent and means ± standard errors. *Robot Action to Person* denotes the model's classification of the robot's action relative to a human (e.g., *yielding to*, *overtaking*, *avoiding*, *following*, *not considering*).

Table 4: **Environment summary across VLMs.** *Overall FR* is the model's failure rate with standard error in smaller type. Environment cells show *failure rate increase* (FRI) relative to the model's overall FR, with raw FR ± SE in smaller parentheses. FRI $= 1$ equals overall; $> 1$ is worse-than-average; $< 1$ is better-than-average for that model.

| Model | Overall FR | Indoors | Outdoors | Blind corner |
|---|---|---|---|---|
| o4-mini | 6.42% ± 0.33% | 0.87× (5.58% ± 0.90%) | 1.02× (6.53% ± 0.36%) | 1.45× (9.28% ± 1.56%) |
| GPT-4o | 23.63% ± 0.60% | 1.39× (32.91% ± 2.15%) | 0.96× (22.65% ± 0.62%) | 1.13× (26.67% ± 2.95%) |
| LLaVa-Next | 33.78% ± 0.67% | 0.91× (30.82% ± 2.11%) | 1.01× (34.09% ± 0.71%) | 0.99× (33.33% ± 3.14%) |
| Gemini 2.0 | 15.00% ± 0.51% | 1.22× (18.24% ± 1.77%) | 0.98× (14.65% ± 0.53%) | 1.16× (17.33% ± 2.52%) |
| Gemini 2.5 | 9.02% ± 0.41% | 0.91× (8.18% ± 1.25%) | 1.01× (9.11% ± 0.43%) | 1.13× (10.22% ± 2.02%) |

Table 5: **Robot Action to Person.** Cells report failure rate (FR); parentheses show the occurrence (%) of the VLM choosing that action. "—" denotes the action was never chosen.

| Model | Yielding to | Overtaking | Following | Avoiding | Not Considering |
|---|---|---|---|---|---|
| o4-mini | 57.14% (3.2%) | 50.00% (1.8%) | 31.25% (3.6%) | 20.59% (7.7%) | 2.70% (83.7%) |
| GPT-4o | 84.85% (8.3%) | 80.00% (1.3%) | 20.00% (1.3%) | 48.84% (53.9%) | 1.42% (35.3%) |
| LLaVa-Next | — | — | — | 53.13% (88.2%) | 6.38% (11.8%) |
| Gemini 2.0 | 66.67% (0.8%) | 100.00% (0.5%) | — | 46.62% (33.3%) | 3.83% (65.4%) |
| Gemini 2.5 | 62.50% (2.0%) | 47.37% (9.5%) | 0.00% (2.8%) | 15.56% (33.8%) | 0.97% (51.9%) |

**GPT-4o.** Overall 24% failure rate; higher indoors than outdoors (32.9% vs 22.6%) and at blind corners (26.7% vs 23.5%); for Robot Action to Person: *yielding to* 85%, *overtaking* 80%, *avoiding* 48.8%, *not considering* 1.4%.

**o4-mini.** Overall 6.4% failure rate; higher at blind corners (9.3% vs 6.2%); Robot Action to Person: *yielding to* 57%, *overtaking* 50%, *following* 31%, *avoiding* 21%, *not considering* 2.7%.

**LLaVa-Next.** Overall 33.8% failure rate; very limited Robot Action to Person diversity—only *avoiding* (53.1%) and *not considering* (6.38%).

**Gemini 2.0.** Overall 15.0% failure rate; higher at blind corners (17.3% vs 14.9%); failure cases show more people in scene (12.26 vs 11.91); Robot Action to Person: *avoiding* 46.6%, *not considering* 3.83%, with rare but often wrong *yielding to* (66.7%) and *overtaking* (100%).

**Gemini 2.5.** Overall 9.0% failure rate; Robot Action to Person shows higher action diversity but some labels remain difficult: *yielding to* 62.5%, *overtaking* 47.4%, versus *avoiding* 15.6%, *not considering* 1.0%, *following* 0%.

### 7.7 Survey Question Details

Here we show the details and qualitative descriptions of questions used throughout the benchmark by providing a question for each VQA prompt, shown in Table 6. We categorize these questions according to their reasoning capability.

Table 6: **Qualitative descriptions of the text components for questions used in SocialNav-SUB**, their pertaining primary reasoning capability, and the number of unique questions through SocialNav-SUB. All questions are multiple-choice questions, with each VQA prompt providing the possible answers. An example of a VQA prompt can be found in Figure 2 and a full example can be found in Appendix 7.5.

| **VLM Reasoning Capability** | **Qualitative Description of Question** | **# of Questions** |
|---|---|---|
| **Spatial** | **Person Initial Position**: The position of the person at the beginning of the video. | 399 |
| | **Person Ending Position**: The position of the person at the end of the video. | 399 |
| | **Goal Initial Position**: The initial position of the goal with respect to the robot's view. | 60 |
| | **Goal End Position**: The end position of the goal with respect to the robot's view. | 60 |
| | **Person End Goal Obstruction**: Whether the person is obstructing the robot's path towards the goal at the end of the video. | 399 |
| **Spatiotemporal** | **Robot Moving Direction**: The direction the robot is moving in the video. | 60 |
| | **Person Distance Change**: The relative distance change of the person to the robot from the beginning of the video to the end. | 399 |
| | **Person Goal Obstruction**: Whether the person is obstructing the robot's path towards the goal during the video. | 399 |
| **Social** | **Robot Affected by Person**: Whether the robot's (human operator's) actions are affected by the person. | 399 |
| | **Robot Action to Person**: The high-level relational action of the robot with respect to the person (e.g., the robot avoided person 2). | 399 |
| | **Person Affected by Robot**: Whether the robot's (human operator's) actions are affected by the person. | 399 |
| | **Person Action to Robot**: The high-level relational action of the person with respect to the robot (e.g., person 2 avoided the robot). | 399 |
| | **Robot Affected by Person at End**: Whether the robot's (human operator's) actions are affected by the person at the end of the video. | 399 |
| | **Robot Action to Person at End**: The high-level relational action of the robot with respect to the person at the end of the video. | 399 |
| | **Person Action to Robot at End**: The high-level relational action of the person with respect to the robot at the end of the video. | 399 |

Table 7: **Full spatial reasoning questions in SocialNav-SUB, with question type and options.** Here, PERSON can be any labeled person in the scene, e.g. Person 3.

| **Question** | **Type** | **Options** |
|---|---|---|
| In the beginning, {PERSON} is ___ the robot. | Multiple Choice | ahead of; to the left of; to the right of; behind |
| At the end, {PERSON} is ___ the robot. | Multiple Choice | ahead of; to the left of; to the right of; behind |
| In the beginning frame, the goal is ___ of the robot. | Multiple Choice | ahead; to the left; to the right |
| At the end frame, the goal is ___ of the robot. | Multiple Choice | ahead; to the left; to the right |
| At the end frame, is {PERSON}'s position in the way of the robot's path to the goal? | Multiple Choice | yes; no |

Table 8: **Full spatiotemporal reasoning questions in SocialNav-SUB, with question type and options.** Here, PERSON can be any labeled person in the scene, e.g. Person 3. We convert the Multiple Select question into Multiple Choice by taking the power set of all options.

| Question | Type | Options |
|---|---|---|
| The robot is _____ (Select all that apply) | Multiple Select | moving ahead; turning left; turning right |
| At the end, {PERSON} ends up ____ the robot compared to the beginning. | Multiple Choice | closer to; further away from; about the same distance to |
| Is {PERSON}'s path in the way of the robot's path to the goal? | Multiple Choice | yes; no |

Table 9: **Full social reasoning questions in SocialNav-SUB, with question type and options.** Here, PERSON can be any labeled person in the scene, e.g. Person 3.

| Question | Type | Options |
|---|---|---|
| Is the robot's movement affected by {PERSON}? | Multiple Choice | yes; no |
| The robot is most likely ___ {PERSON}. | Multiple Choice | avoiding; overtaking; not considering; following; yielding to |
| Is {PERSON}'s movement affected by the robot? | Multiple Choice | yes; no |
| {PERSON} is most likely ___ the robot. | Multiple Choice | avoiding; overtaking; not considering; following; yielding to |
| In the future (after the end of the video), should the robot's movement towards the goal be affected by {PERSON}? | Multiple Choice | yes; no |
| In the future (after the end of the video), the robot should ___ {PERSON} as it makes progress towards the goal. | Multiple Choice | avoid; overtake; not consider; follow; yield to |
| In the future (after the end of the video), {PERSON} will most likely ___ the robot as the robot attempts to make progress towards the goal. | Multiple Choice | avoid; overtake; not consider; follow; yield to |

### 7.8 Main Experiment Question Results

We provide the question-level performance for the main experiment results from Section 4.2 for the VLMs shown in Table 10, reasoning-based VLMs shown in 11, and the baselines shown in Table 12.

Table 10: **Performance Across Individual Questions for non-reasoning VLMs.** These results highlight the deficiencies of non-reasoning VLMs: 1) Gemini [7] has stronger social reasoning than other non-reasoning VLMs for most questions but has worse spatial reasoning performance across most tasks compared to GPT-4o; 2) LLaVa-Next-Video [17] has poor spatial reasoning performance for most questions, determining the moving direction of the robot, and poor ability to infer the future action of the robot, but performs well for certain questions such as determining whether somebody is obstructing the goal and some social reasoning questions; 3) GPT-4o [6] has moderate performance across tasks but lacks strong social reasoning.

| **Category** | **Question Name** | **Gemini 2.0** | | **GPT-4o** | | **LLaVa-Next-Video** | |
|---|---|---|---|---|---|---|---|
| | | **PA** | **CW PA** | **PA** | **CW PA** | **PA** | **CW PA** |
| **Spatial** | Person Initial Position | 0.52 ± 0.01 | 0.81 ± 0.01 | 0.54 ± 0.01 | 0.84 ± 0.01 | 0.05 ± 0.00 | 0.10 ± 0.01 |
| | Person Ending Position | 0.38 ± 0.01 | 0.64 ± 0.02 | 0.43 ± 0.01 | 0.71 ± 0.02 | 0.24 ± 0.01 | 0.44 ± 0.02 |
| | Goal Initial Position | 0.69 ± 0.04 | 0.85 ± 0.04 | 0.74 ± 0.03 | 0.92 ± 0.03 | 0.14 ± 0.02 | 0.20 ± 0.04 |
| | Goal End Position | 0.56 ± 0.04 | 0.73 ± 0.05 | 0.65 ± 0.04 | 0.83 ± 0.04 | 0.15 ± 0.02 | 0.22 ± 0.04 |
| | Person End Goal Obstruction | 0.74 ± 0.01 | 0.86 ± 0.02 | 0.68 ± 0.02 | 0.78 ± 0.02 | 0.80 ± 0.01 | 0.93 ± 0.01 |
| **Spatiotemporal** | Robot Moving Direction | 0.46 ± 0.05 | 0.64 ± 0.05 | 0.57 ± 0.04 | 0.81 ± 0.04 | 0.24 ± 0.04 | 0.38 ± 0.05 |
| | Person Distance Change | 0.31 ± 0.01 | 0.53 ± 0.02 | 0.46 ± 0.01 | 0.74 ± 0.02 | 0.47 ± 0.01 | 0.75 ± 0.02 |
| | Person Goal Obstruction | 0.62 ± 0.02 | 0.76 ± 0.02 | 0.54 ± 0.02 | 0.67 ± 0.02 | 0.74 ± 0.01 | 0.89 ± 0.01 |
| **Social** | Robot Affected by Person | 0.64 ± 0.02 | 0.78 ± 0.02 | 0.50 ± 0.02 | 0.63 ± 0.02 | 0.75 ± 0.01 | 0.91 ± 0.01 |
| | Robot Action to Person | 0.51 ± 0.01 | 0.75 ± 0.02 | 0.37 ± 0.01 | 0.57 ± 0.02 | 0.25 ± 0.01 | 0.42 ± 0.02 |
| | Person Affected by Robot | 0.74 ± 0.01 | 0.88 ± 0.01 | 0.58 ± 0.02 | 0.71 ± 0.02 | 0.79 ± 0.01 | 0.94 ± 0.01 |
| | Person Action to Robot | 0.62 ± 0.01 | 0.86 ± 0.02 | 0.45 ± 0.02 | 0.65 ± 0.02 | 0.67 ± 0.01 | 0.92 ± 0.01 |
| | Robot Affected by Person at end | 0.72 ± 0.01 | 0.87 ± 0.01 | 0.55 ± 0.02 | 0.68 ± 0.02 | 0.79 ± 0.01 | 0.94 ± 0.01 |
| | Robot Action to Person at end | 0.60 ± 0.01 | 0.85 ± 0.02 | 0.41 ± 0.02 | 0.59 ± 0.02 | 0.08 ± 0.01 | 0.14 ± 0.01 |
| | Person Action to Robot at end | 0.62 ± 0.01 | 0.87 ± 0.01 | 0.40 ± 0.02 | 0.59 ± 0.02 | 0.03 ± 0.00 | 0.05 ± 0.01 |

Table 11: **Performance Across Individual Questions for Large Reasoning Models.** These results indicate that o4-mini displays worse performance across most spatial reasoning question but has strong performance on determining if a person is obstructing the path to the goal. We hypothesize, with evidence in Appendix 7.1, that better performance in these questions can result in better social reasoning performance and may be a limiting factor for o4-mini. Gemini 2.5 shows worse performance across spatiotemporal reasoning and social reasoning compared to o4-mini but comparable performance in spatial reasoning. Gemini 2.5 has a particularly difficult time in determining the moving direction of the robot compared to other models. Although we evaluated using o4-mini and Gemini 2.5 flash, we expect that these may be lower bounds on the performance for their higher-end model variations.

| **Category** | **Question Name** | **Gemini 2.5** | | **o4-mini** | |
|---|---|---|---|---|---|
| | | **PA** | **CW PA** | **PA** | **CW PA** |
| **Spatial** | Person Initial Position | 0.49 ± 0.01 | 0.78 ± 0.02 | 0.49 ± 0.01 | 0.76 ± 0.02 |
| | Person Ending Position | 0.36 ± 0.01 | 0.59 ± 0.02 | 0.36 ± 0.01 | 0.58 ± 0.02 |
| | Goal Initial Position | 0.70 ± 0.04 | 0.86 ± 0.04 | 0.48 ± 0.05 | 0.58 ± 0.06 |
| | Goal End Position | 0.52 ± 0.04 | 0.67 ± 0.05 | 0.48 ± 0.05 | 0.60 ± 0.06 |
| | Person End Goal Obstruction | 0.66 ± 0.02 | 0.77 ± 0.02 | 0.81 ± 0.01 | 0.93 ± 0.01 |
| **Spatiotemporal** | Robot Moving Direction | 0.34 ± 0.04 | 0.50 ± 0.06 | 0.56 ± 0.04 | 0.80 ± 0.04 |
| | Person Distance Change | 0.47 ± 0.01 | 0.75 ± 0.02 | 0.45 ± 0.01 | 0.71 ± 0.02 |
| | Person Goal Obstruction | 0.60 ± 0.02 | 0.73 ± 0.02 | 0.73 ± 0.01 | 0.87 ± 0.01 |
| **Social** | Robot Affected by Person | 0.57 ± 0.02 | 0.71 ± 0.02 | 0.73 ± 0.01 | 0.89 ± 0.01 |
| | Robot Action to Person | 0.44 ± 0.02 | 0.67 ± 0.02 | 0.58 ± 0.01 | 0.84 ± 0.02 |
| | Person Affected by Robot | 0.70 ± 0.02 | 0.83 ± 0.02 | 0.77 ± 0.01 | 0.91 ± 0.01 |
| | Person Action to Robot | 0.58 ± 0.02 | 0.80 ± 0.02 | 0.60 ± 0.02 | 0.83 ± 0.02 |
| | Robot Affected by Person at end | 0.56 ± 0.02 | 0.68 ± 0.02 | 0.77 ± 0.01 | 0.91 ± 0.01 |
| | Robot Action to Person at end | 0.44 ± 0.02 | 0.62 ± 0.02 | 0.62 ± 0.01 | 0.87 ± 0.01 |
| | Person Action to Robot at end | 0.58 ± 0.01 | 0.81 ± 0.02 | 0.58 ± 0.02 | 0.82 ± 0.02 |

Table 12: **Performance Across Individual Questions for Baselines.** For the Human Oracle and Average Human baselines, these results highlight questions that humans disagreed on more often, showing that determining spatial labels for humans was more disagreeable than social reasoning questions. The rule-based baseline performance indicates that it struggles with determining what the initial and ending position of humans are as well as determining if a person gets closer to further away, showing that it is not as trivial as determining a cutoff value for this from rules described in 7.10.

| **Category** | **Question Name** | **Human Oracle** | | **Average Human** | | **Rule-Based** | |
|---|---|---|---|---|---|---|---|
| | | **PA** | **CW PA** | **PA** | **CW PA** | **PA** | **CW PA** |
| **Spatial** | Person Initial Position | 0.64 ± 0.01 | 1.00 ± 0.00 | 0.46 ± 0.01 | 0.73 ± 0.00 | 0.49 ± 0.01 | 0.78 ± 0.01 |
| | Person Ending Position | 0.61 ± 0.01 | 1.00 ± 0.00 | 0.43 ± 0.01 | 0.71 ± 0.01 | 0.41 ± 0.01 | 0.67 ± 0.02 |
| | Goal Initial Position | 0.80 ± 0.02 | 1.00 ± 0.00 | 0.68 ± 0.03 | 0.85 ± 0.01 | 0.68 ± 0.04 | 0.83 ± 0.05 |
| | Goal End Position | 0.77 ± 0.02 | 1.00 ± 0.00 | 0.62 ± 0.02 | 0.82 ± 0.01 | 0.56 ± 0.04 | 0.72 ± 0.05 |
| | Person End Goal Obstruction | 0.86 ± 0.01 | 1.00 ± 0.00 | 0.77 ± 0.01 | 0.89 ± 0.00 | 0.80 ± 0.01 | 0.93 ± 0.01 |
| **Spatiotemporal** | Robot Moving Direction | 0.69 ± 0.03 | 1.00 ± 0.00 | 0.52 ± 0.03 | 0.74 ± 0.02 | 0.62 ± 0.04 | 0.87 ± 0.03 |
| | Person Distance Change | 0.63 ± 0.01 | 1.00 ± 0.00 | 0.46 ± 0.01 | 0.74 ± 0.00 | 0.48 ± 0.01 | 0.76 ± 0.02 |
| | Person Goal Obstruction | 0.83 ± 0.01 | 1.00 ± 0.00 | 0.73 ± 0.01 | 0.88 ± 0.00 | 0.78 ± 0.01 | 0.94 ± 0.01 |
| **Social** | Robot Affected by Person | 0.82 ± 0.01 | 1.00 ± 0.00 | 0.72 ± 0.01 | 0.87 ± 0.00 | 0.76 ± 0.01 | 0.91 ± 0.01 |
| | Robot Action to Person | 0.67 ± 0.01 | 1.00 ± 0.00 | 0.50 ± 0.01 | 0.74 ± 0.01 | 0.57 ± 0.01 | 0.82 ± 0.02 |
| | Person Affected by Robot | 0.84 ± 0.01 | 1.00 ± 0.00 | 0.74 ± 0.01 | 0.88 ± 0.00 | 0.79 ± 0.01 | 0.94 ± 0.01 |
| | Person Action to Robot | 0.72 ± 0.01 | 1.00 ± 0.00 | 0.56 ± 0.01 | 0.77 ± 0.01 | 0.67 ± 0.01 | 0.93 ± 0.01 |
| | Robot Affected by Person at end | 0.84 ± 0.01 | 1.00 ± 0.00 | 0.73 ± 0.01 | 0.88 ± 0.00 | 0.79 ± 0.01 | 0.94 ± 0.01 |
| | Robot Action to Person at end | 0.70 ± 0.01 | 1.00 ± 0.00 | 0.53 ± 0.01 | 0.75 ± 0.01 | 0.67 ± 0.01 | 0.94 ± 0.01 |
| | Person Action to Robot at end | 0.71 ± 0.01 | 1.00 ± 0.00 | 0.54 ± 0.01 | 0.76 ± 0.01 | 0.70 ± 0.01 | 0.98 ± 0.01 |

### 7.9 Ablation Experiments

Table 13: **Ablation experiment of querying strategies**. The metric used is Probability of Agreement (PA). The baseline row BEV+CoT represents the VLM's performance with both CoT and BEV prompts enabled, while the subsequent rows show the effects of removing either CoT or BEV components.

| **Model** | **Ablation** | **Spatial Reasoning** | **Spatiotemporal Reasoning** | **Social Reasoning** |
|---|---|---|---|---|
| GPT-4o | CoT+BEV | 0.56 ± 0.01 | 0.51 ± 0.01 | 0.47 ± 0.01 |
| | No CoT | 0.58 ± 0.01 | 0.53 ± 0.01 | 0.35 ± 0.01 |
| | No BEV | 0.51 ± 0.01 | 0.44 ± 0.01 | 0.42 ± 0.01 |
| LLaVa-Next-Video | CoT+BEV | 0.35 ± 0.01 | 0.58 ± 0.01 | 0.48 ± 0.01 |
| | No CoT | 0.35 ± 0.01 | 0.58 ± 0.01 | 0.38 ± 0.01 |
| | No BEV | 0.35 ± 0.01 | 0.61 ± 0.01 | 0.46 ± 0.01 |
| Gemini 2.0 | CoT+BEV | 0.55 ± 0.01 | 0.46 ± 0.01 | 0.63 ± 0.01 |
| | No CoT | 0.56 ± 0.01 | 0.48 ± 0.01 | 0.58 ± 0.01 |
| | No BEV | 0.56 ± 0.01 | 0.46 ± 0.01 | 0.64 ± 0.01 |

To understand the impact of specific querying strategies on model performance, we conducted ablation experiments, systematically removing components such as chain-of-thought (CoT) reasoning and BEV prompts. Table 13 summarizes how these ablations affect PA in spatial, spatio-temporal, and social reasoning tasks.

**CoT reasoning.** The results indicate that removing the CoT component does not significantly affect spatial and spatiotemporal reasoning performance. However, the removal of CoT leads to a notable decrease in social reasoning performance across all models. We hypothesize that social reasoning tasks more often require multi-step reasoning to which CoT can help structure complex chains of inference.

**BEV visual prompts.** The results from removing BEV prompts indicate that there is not a significant effect across the capabilities for LLaVa-Next-Video and Gemini 2.0, but provides a notable decrease in performance for GPT-4o across all capabilities. While these results may not indicate a clear winner for all models, it suggests that prompt design remains an open question which needs to be further studied, an endeavor that can be pursued using our benchmark.

Table 14: **Gemini ablation experiments when using ground truth spatial and spatiotemporal answers for CoT reasoning.** Our results indicate that better spatial reasoning and spatiotemporal reasoning leads to better performance on social reasoning questions.

| **Question Name** | **CoT** | | **CoT with Ground-Truth Spatial(Temporal) Reasoning** | |
|---|---|---|---|---|
| | **PA** | **CW PA** | **PA** | **CW PA** |
| Robot Affected by Person | 0.64 ± 0.02 | 0.78 ± 0.02 | 0.78 ± 0.01 | 0.94 ± 0.01 |
| Robot Action to Person | 0.51 ± 0.01 | 0.75 ± 0.02 | 0.60 ± 0.01 | 0.88 ± 0.01 |
| Person Affected by Robot | 0.74 ± 0.01 | 0.88 ± 0.01 | 0.78 ± 0.01 | 0.94 ± 0.01 |
| Person Action to Robot | 0.62 ± 0.01 | 0.86 ± 0.02 | 0.65 ± 0.01 | 0.90 ± 0.01 |
| Robot Affected by Person at end | 0.72 ± 0.01 | 0.87 ± 0.01 | 0.78 ± 0.01 | 0.93 ± 0.01 |
| Robot Action to Person at end | 0.60 ± 0.01 | 0.85 ± 0.02 | 0.65 ± 0.01 | 0.91 ± 0.01 |
| Person Action to Robot at end | 0.62 ± 0.01 | 0.87 ± 0.01 | 0.65 ± 0.01 | 0.91 ± 0.01 |

**Spatial Reasoning's Affect on Performance.** We ran an additional experiment to see if a lack of strong performance for spatial and spatiotemporal reasoning was affecting performance on social reasoning questions. Table 14 shows the results of running this experiment where we used the human consensus answer's for the answers for spatial and spatiotemporal reasoning questions for the VLM, which was also provided as chain-of-thought reasoning to the VLM in the form of context; the VLM was then evaluated on social reasoning questions. These results indicate that a strong spatial and spatiotemporal reasoning capabilities can lead to significantly better performance on social reasoning questions. The "Person Goal Obstruction" question may provide sufficient information for the VLM to easily answer the "Robot Affected By Person" question, to which we run an additional experiment and empirically found that, although it was not as drastic, there were performance gains across all questions. These results indicate that hybrid VLM systems that help VLM's with their weaknesses (such as dedicated perception modules) may be more effective rather than entirely relying on the VLM for all questions.

### 7.10 Rule-Based Baseline Details

As mentioned in Section 4.1, we developed a rule-based baseline which uses a set of hand-crafted rules to determine answers for VQA questions. Although our simple approach demonstrates better performance than VLMs, it is by no means comprehensive and more complex rules can be devised to further push performance. We briefly summarize the simple rules to determine answers for our Rule-Baed baseline:

- Spatial Reasoning Position Questions: Determine deviation in the horizontal direction and use it along with cutoff values to determine whether to answer they are to the left, ahead, or behind.
- Goal Obstruction Questions: Draw a line from the robot to the goal and a line from the person's trajectory, if the lines intersect, consider the person obstructing the goal.
- Person Distance Change: Look at the initial relative position and end relative positions for the person, determine the appropriate answer based on the distance between the two points.
- Robot Moving Direction: Use the horizontal deviation between the the initial relative position of the robot and the end relative position to determine if the robot is turning.
- Social Reasoning "Affected" Questions: If the person is obstructing the goal, then answer that the robot will be affected by them.
- Social Reasoning "Action" Questions: If the person is obstructing the goal, then avoid the person. For person action questions, use the same answer as the robot action questions.